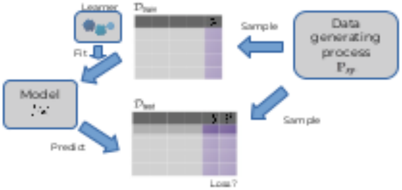
GE FOR A FIXED MODEL

- GE for a fixed model: $\mathrm{GE}\left(\hat{f}, L\right) := \mathbb{E}\left[L\left(y, \hat{f}(\mathbf{x})\right)\right]$
  Expectation over a single, random test point $(\mathbf{x}, y) \sim \mathbb{P}_{xy}$.
- Estimator, **if a dedicated test set is available** (size m)

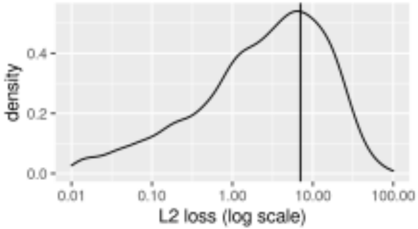
$$\widehat{\mathrm{GE}}(\hat{f}, L) := \frac{1}{m} \sum_{(\mathbf{x}, y) \in \mathcal{D}_{\text{test}}} \left[L\left(y, \hat{f}(\mathbf{x})\right)\right]$$



NB: Very often, no dedicated test-set is available, and what we describe here is not same as hold-out splitting (see later).

EXAMPLE: TEST LOSS AS RANDOM VARIABLE

- For a fixed model and dedicated i.i.d. test set, we can easily approximate the complete test loss distribution $L(y, \hat{f}(\mathbf{x}))$.
- LM on `mlbench::friedman1` test problem
- With $n_{\text{train}} = 500$ we create a fixed model
- We feed 5000 fresh test points to model
- And plot the pointwise $L2$ loss.



- The result is a unimodal distribution with long tails.
- Mean and one standard deviation to either side are highlighted in grey.

GENERALIZATION ERROR FOR INDUCER

$$\mathrm{GE}(\mathcal{I}, \boldsymbol{\lambda}, n_{\mathrm{train}}, \rho) := \lim_{n_{\mathrm{test}} \to \infty} \mathbb{E}\left[\rho\left(\mathbf{y}, \boldsymbol{F}_{\mathcal{D}_{\mathrm{test}}, \mathcal{I}(\mathcal{D}_{\mathrm{train}}, \boldsymbol{\lambda})}\right)\right]$$

- Quality of models when fitted with $\mathcal{I}_{\boldsymbol{\lambda}}$ on n_{train} points from $\mathbb{P}_{xy}$.
- Expectation **both** over $\mathcal{D}_{\mathrm{train}}$ and $\mathcal{D}_{\mathrm{test}}$, sampled independently.
- This is estimated by all following **resampling** procedures.
- NB: All of the models produced during that phase of evaluation are only intermediate results.

GENERALIZATION ERROR FOR INDUCER

$$\mathrm{GE}(\mathcal{I}, \boldsymbol{\lambda}, n_{\mathrm{train}}, \rho) := \lim_{n_{\mathrm{test}} \to \infty} \mathbb{E}\left[\rho\left(\mathbf{y}, \boldsymbol{F}_{\mathcal{D}_{\mathrm{test}}, (\mathcal{I}(\mathcal{D}_{\mathrm{train}}, \boldsymbol{\lambda}))}\right)\right]$$

- We can already see a potential source of pessimistic bias in our estimator: While we would like to estimate a GE with $n_{\mathrm{train}} = |\mathcal{D}|$, the size of the complete data set, in practice we can only do this for strictly smaller values, so that test data is left to work with.
- For pointwise losses ρ_L:

$$\mathrm{GE}(\mathcal{I}, \boldsymbol{\lambda}, n_{\mathrm{train}}, \rho_L) := \mathbb{E}\left[L(y, \mathcal{I}(\mathcal{D}_{\mathrm{train}}, \boldsymbol{\lambda})(\mathbf{x}))\right]$$

Expectation **both** over $\mathcal{D}_{\mathrm{train}}$ and $(\mathbf{x}, y)$ independently from $\mathbb{P}_{xy}$.

- Retcon for GE of model: GE of learner, conditional on $\mathcal{D}_{\mathrm{train}}$

$$\mathrm{GE}\left(\hat{f}, L\right) := \mathrm{GE}(\mathcal{I}, \boldsymbol{\lambda}, n_{\mathrm{train}}, \rho_L | \mathcal{D}_{\mathrm{train}})$$

if $\hat{f} = \mathcal{I}(\mathcal{D}_{\mathrm{train}}, \boldsymbol{\lambda})$ and $n_{\mathrm{train}} = |\mathcal{D}_{\mathrm{train}}|$.